Using Weights in the Analysis of Survey Data

What do we mean by Survey weights?

- A Value assigned to each case in the data file
- Usually it used to make statistics computed from the data more representative of the population benchmark
- E.g. A value shall indicate how much each case will count in a statistical procedure
- Examples:
  -- A weight of 4 means that a case is counted as four in the dataset
  -- A weight of 1 means that a case is counted as one in the dataset
  -- Weights can (and often are) fractions, but are always positive and non-zero
- In stats we call them as *pweight*

Types of Survey Weights

Two most common type:

-- Design Weights
-- Post-Stratification or Non-responsive weights

Design Weight:
-- Normally used to compensate for over – or under – sampling of specific cases or for disproportionate stratification

--Example: it is common practice to over-sample minority group members or persons living in areas with larger percentage minority. If we doubled the size of our sample from minority areas, then each case in that area would get a design weight of ½ or 0.5

-- The design weight when we want the statistics to be representative of the population

Post-Stratification Weights

Post stratification or Non responsive weight

-- This type is used to compensate for the fact that persons with certain characteristics are not as likely to respond to the survey.

-- *Example.* Most general population surveys have substantially more female than male respondents (often 70/30) even though there are more males in the population. Now survey over represents female and under represents male in the population a weight is assigned to compensate for this bias.

-- There are many respondents characteristics that are more likely to be related to the propensity to respond

- Age
- Education
- Race/Ethnicity
- Gender
- Place of residence

Post-Stratification Weights

Post stratification or Non responsive weight

-- This type is used to compensate for the fact that persons with certain characteristics are not as likely to respond to the survey.

-- *Example.* Most general population surveys have substantially more female than male respondents (often 70/30) even though there are more males in the population. Now survey over represents female and under represents male in the population a weight is assigned to compensate for this bias.

-- There are many respondents characteristics that are more likely to be related to the propensity to respond

- Age
- Education
- Race/Ethnicity
- Gender
- Place of residence

How Do We Calculate Weights

For Analysis, Only one weight per case can be used. If we weight for different factors, these weights must be combined together into one weight.

Ex. We have a design weight (Dweight) and a post stratification (pweight) for each case.

To calculate a Total weight these are multiplied together
Total Weight = Dweight * Pweight

Note: We never give a weight the value of zero unless we want the case excluded from our analysis. It's default value should be 1.

Calculating Design Weights

If we know the sampling fraction for each case, the weight is the inverse of the sampling fraction.

Design Weight = 1/(Sampling fraction)

The sampling fraction could also be the over sampling amount for a given group or area

Ex. If we over sampled African Americans at a rate 4 times greater then the rate of whites, then the design weight for an frican American would be ¼ and for a white respondent would be 1.

Calculating Post Stratification Weights

This is normally more difficult then design weights

It requires the use of auxiliary information about the population and may take a number of different variables into account

Information usually needed:
-- Population estimates of the distribution of a set of demographic characteristics that have also been measured in the sample
-- For Example, information found in the census such as:

- Gender
- Age
- Educational Attainment
- Household size
- Residence (e.g., rural, urban, metropolitan)
- Region

Calculating Post Stratification Weights

Gender	Population Proportion	Sample proportion	Population/ Sample	Weight
Female	.5	.7	.5/.7	.714
Male	.5	.4	.5/.4	1.25
Total	1	1		

Census report is used to find the gender distribution in the population (50% female). This is compared to the gender distribution in the sample of completed interviews (70 female).

Problem: What if we have more than one characteristic to balance with the population?

Automatic Iterative Solutions

A procedure called Raking has been programmed by several folks. Is relatively widely used.
Our team has a SAS Raking macro which automates the iterative task.
In the SAS macro we can set several options, such as how accurate you want to weight, and also can impose some limits on the size of weights (min and max)

Data Analysis methods with Weighted Data

-- Should use a statistical procedure that adjusts for the impact of the weights on the standard error. Standard errors based on the actual N and not the weighted N.

-- Normalization of weights
Setting the weights so that the N in the weighted data equals N in the unweighted data.
To calculate, multiply the weight by (Unweighted N / Weighted N)
If the statistical procedure does not use weights correctly for the standard errors, normalization is a less biased choice.

-- Another choice is to not use weights at all for regression models, Instead include all the variables used to create the weights as independent variables. Results in unbiased estimates and standard errors.

What Weights to use in Analysis of Longitudinal (Panel) data?

Many panel datasets have several weights to choose among.
-- Cross sectional weights.
-- Weights for each panel if multiple panels

Weights to use will primarily depend on the data analysis methods used.

Longitudinal Panel weights are usually computed from two components
-- The cross sectional weight from the previous panel or the first panel.
-- A weight calculated to adjust for attrition.

Calculating the non- response (attrition) weight component
-- usually use logistic regression with response to the wave as outcome variable (O = no; 1 = yes)
Predict probability of responding
Inverse of this probability is the attrition weight.

When to use Unweighted Data

If the sample is not self-weighted then it is a good idea to use weights as often as possible

Some methods don't allow weights E.g., some multilevel models, some structural equation programs etc

Steps to follow to avoid bias in unweighted analyses
--Include s independent variables in the models all the variables that might account for the disproportionate sample design or non-response
-- If a weight is available, the weight itself could be included as an independent variable
-- If the weight has significant effect on the outcome of the model including design variables, then it suggest that the weight has been constructed in a way related to the dependent variable. A bias is possible
-- Compare weighted and unweighted results from the methods that allow weights. If substantive differences, then weight yield a bias.
-- Weighting has a larger effect on descriptive statistics then on regression coefficients.